doi:10.5128/ERYa6.11

EESTI RAKENDUSLINGVISTIKA ÜHINGU AASTARAAMAT 6, 175–193

KESKSETE LAUSEKOMPONENTIDE JÄRJESTUS ÕPPIJAKEELES: ARVUTIANALÜÜSI KATSE

Helena Metslang, Erika Matsak

Ülevaade. Artikkel käsitleb eesti keele lihtlause sõnajärje arvutianalüüsi katset, mille eesmärgiks on õppijakeele sõnajärje vealeidja loomine. Katse käigus koostati eesti keele sagedaste sõnajärjetüüpide mallid, mis kirjeldasid lihtlause ja mõne lihtsama liitlausetüübi verbi, tuumargumentide ning nende järge mõjutavate moodustajate või sõnade järge (põhiliselt subjekt, objekt, predikaat, adverbiaal lause algul või seotud laiendina, üldlaiend). Mallid leiti Tartu Ülikooli kirjakeele korpuse põhjal. Saadud mallide katvust hinnati kirjakeele ja õppijakeele korpuste peal spetsiaalselt loodud programmi abil. Artiklis kirjeldatav programm, mis on kasutatav koos mallide koguga, analüüsib õppijakeelt, märkides küsitavaks laused, mis ühelegi mallile ei vasta. Artikkel tutvustab mallide kogu loomise protsessi ja tekstilausete sõnajärge hindavat programmi. Antakse ka ülevaade programmi efektiivsusest õppijakeele tekstide analüüsil ning vealeidja edasise arendamise vajadustest. Õppijakeele analüüsil kasutati Tallinna Ülikooli eesti vahekeele korpust, mis koondab ligi 740 000 sõne mahus eesti keele õppijate loovkirjutisi ja harjutusi.*

Võtmesõnad: sõnajärg, korpuslingvistika, teise keele omandamine, eesti keel

1. Sissejuhatus

Teise keele õppimine on keerukas protsess ja õppija peab olema vapper ega tohi karta emakeelekõnelejate muigvel suid, kui ta on taas mõne iseäranis kohmaka lause koostamisega hakkama saanud. Artikli allikmaterjalis leidus neid rohkesti, näiteks *Ühel hetkel mul tundakse, et olin Vana Egiptus*, *Pärast puhkemist ma uuesti sõitsin TTÜsse*.

* Tööd on toetanud riiklik programm "Eesti keele keeletehnoloogiline tugi (2006–2010)", projekt R0807 "VAKO – Eesti vahekeele korpuse keeletarkvara ja keeletehnoloogilise ressursi arendamine" ning sihtteema SF0180084s08 "Eesti keele morfosüntaktiline ehitus ja areng". Täname Vahur Rebast, kes aitas vealeidja prototüüpi programmeerida, Kaili Müüriseppa hea nõu eest morfosüntaktilise arvutianalüüsi küsimustes, Raili Pooli, kes võimaldas meile ligipääsu oma rektsioonisõnastiku (Pool 1999) elektroonilisele versioonile, ning anonüümseid retsensente kasulike tähelepanekute eest.

Üks keerukaid valdkondi keeleõppijale on sõnajärg. Käsitleme siin õppijakeele tuumlause sõnajärje automaatse tuvastamise küsimust, mis omakorda on tihedalt seotud kohustuslike lauseelementide õige määramisega. Keelespetsialisti ja informaatiku koostöös sündinud artikkel kirjeldab õppijakeele sõnajärge, tuumlause ehituse probleeme ning meetodit, kuidas oleme neid nähtusi automaatselt, süntaksist lähtuvalt analüüsinud. Kuna lause sõnajärg on kompleksne ning mitte täiesti reeglistatav keelevaldkond, siis kirjeldame siinkohal vaid esimesi samme nende ülesannete lahendamisel. Artikli sissejuhatav osa annab ülevaate olulisematest eesti sõnajärjeuuringutest ning meie lingvistilistest lähtekohtadest ja eesmärkidest. Seejärel iseloomustame Tallinna Ülikooli eesti vahekeele korpuse (edaspidi EVKK)[1] materjali põhjal õppijakeele lausestruktuuri raskuskohti. Järgnevas osas tutvustame sõnajärje vealeidjat ja mallide kogu ning nende loomise käiku. Artikli lõpuosas hindame vealeidja töö edukust, seda mõjutavaid tegureid ja kirjeldame programmi töö tulemuste parandamise võimalusi.

2. Eesti keele sõnajärjest

Sõnajärge võib vaadelda tasanditi. Fraasitasandi sõnajärjes mängib rolli fraasituuma ja selle laiendite järjestus (nt nimisõnafraasis adjektiivifraaside ning neis omakorda adjektiivi laiendavate adverbide järjestus). Lihtlauses on lausetasandi sõnajärje puhul oluline subjekti, predikaadi (või mitmesõnalise predikaadi osade), seotud ja vabade laiendite ning ka muude tekstilause sõnajärge mõjutavate lauseelementide paiknemine (näiteks üldlaiendid). Kuigi paljudes kontekstides on eesti lausetasandi sõnajärg üsna vaba, kehtib V2-reegel enamasti nii normaal-, olemasolu- kui kogeja-omajalauses (Helle Metslang jt 2003: 92–93). Eesti keele sõnajärjemallide sagedusi on vaadelnud Kaja Tael (1988) ja leidnud, et eesti lausete sõnajärjes on domineerivateks kombinatsioonideks subjekt-verb-X (25%), X-verb-subjekt (10%, verb asub sundpositsioonil, verbi muu asukoht muudaks tekstilause vastuvõetamatuks), X-verb-X (10%), verb-X (10%), X-verb-subjekt-X (9%). X tähistab kõigis mallides verbi laiendit, lause kõrvalliiget, mida võib olla üks või enam. (Tael 1988: 3–5)

Tael on tekstilause sõnajärje analüüsil lähtunud järgmistest näitajatest:

- fraaside süntaktilised omadused (laiendite liik ja asukoht, sõnade arv tarindis);
- fraaside semantilised omadused (fraasi semantiline roll ja semantiline tüüp);
- fraaside infostruktuurilised omadused (nt fraasi kuulumine tugevasse või nõrka teemasse või reemasse ning fraasi poolt kantud info olulisus).

Oma töös on ta kasutanud tervelt 67 sõnajärje varieerumisega seotud muutujat (nt *subjekt*, *topikalisatsiooni põhjus*) ning neist igaühe all on 0–15 alakategooriat (nt *elusolend*, *emfaas*, *kontrast*). (Tael 1988: 3, 47)

Eesti sõnajärje aluseks on pigem infostruktuurilised kui süntaktilised printsiibid, eriti suulise kõne puhul (Lindström 2005: 173). Nii võib see olla ka õppijakeele puhul, kes enamasti kuni üsna kõrge keeleoskustasemeni ei erista hästi eri registritest pärit sisendit ning kelle kirjalike tekstide sõnajärg võib olla suulisest

[1] evkk.tlu.ee (10.02.2010).

kõnest mõjutatud. Siiski leiame, et mitmekesiste süntaktilise analüüsi võimaluste olemasolul[2] tasub uurida, kui palju toetab süntaktiline info sõnajärje ja tuumlause argumentide uurimist.

Põhilised eesti sõnajärje tendentsid on süstematiseerinud Nikolai Remmel (1963), tuues välja neutraalse lause erinevate liikmete paiknemise seaduspärad (vt lähemalt jaotis 4). Ta sõnastas lause terviklikkuse seaduse, mis aitab edukalt selgitada ka õppijakeele sõnajärjeküsitavusi:

> .. sõnal, mis lausele terviklikkuse (või lõpliku terviklikkuse) annab, on tendents pealause lõppu asetuda, kuna aga öeldise laiend, mida kõige hõlpsamini saab lausest välja jätta, püüab, vastupidi, vahetult öeldise järele paigutuda .. (Remmel 1963: 260–261)

Remmeli (1963: 259) järgi annab rahulikule neutraalsele lausele terviklikkuse näiteks sihitis, kui see laiendab sihilist verbi, mis ei vaja peale sihitise muud laiendit, nt *Ta lõpetas eile töö*. Valdavalt deskriptiivne käsitlus annab ülevaate kesksete lauseliikmete paiknemisest pealauses, abimäärsõnade ja verbi infiniitsete vormide paigutusest ning ka sellistest spetsiifilisematest kontekstidest nagu kõrvallaused, küsilaused, emotsionaalsed ja erirõhuga laused.

Meie vaadeldava lausetüübi ja süntaksianalüüsiga (vt lähemalt jaotis 3) seostuvad ka järeldused, mille Liina Lindström on teinud suulise kõne sõnajärjest. Kui lause alguses on sihtis, öeldistäide, põhjus- või viisimäärus või kui subjektiks on pronoomen (peamiselt esimesele või teisele isikule viitav), siis on lauses sagedamini subjekti ja verbi otsejärg. Pöördjärjestust eelistatakse juhtudel, kui lause algab valdajamäärusega või kui subjektiks on täisnimisõnafraas. (Lindström 2005: 173–174) Eestikeelsete ilukirjandustekstide sõnajärje kohta on teada, et verbi ees kaldub verbifraasis paiknema adverbiaal – objekt ja predikatiiv on seal harvem. Predikatiivilausetes on sõnajärg tavaliselt SVX, muudes *olema*-verbiga lausetes XVS (eksistentsiaallaused ja kogeja-omajalaused). (Huumo 1994: 280–285) Osalausete paiknemisest eesti liitlauses on ilmunud doktoriväitekiri Kirsi Höglundilt (2006).

3. Sõnajärje arvutianalüüsi eesmärgid

Kavandasime esmaseks uurimisobjektiks lihtlause (või ka võimalikult neutraalse osalause), mis oleks võimalikult terviklik ning mille tuumargumentide struktuur ja sõnajärg oleksid vähe mõjutatud teisestest grammatilistest teguritest. Näiteks ei soovinud me uurimusse kaasata põimlause kõrvallauset, sest sellest võib kontekstiellipsi tõttu puududa mis tahes komponent, ka predikaat, selle verbivorm võib olla infiniitne ning sõnajärg võib olla eripärane (EKG II: 276). Samas soovisime, et laused oleksid tekstilaused, sest üldjuhul tegelevad õppijad just kontekstisidusate lausete harjutamisega. Lisaks täislausetele olid siiski vaatluse all ka narratiivses tekstis laialt levinud aluseta lünklaused. Tähelepanu alt jäid välja umbisikulises tegumoes laused ning hüüd- ja küsilaused ning võimalusel ka mitmesõnalisi predikaate sisaldavad laused.

Ennekõike pakkus huvi lause struktuurne kese: predikaat ja ta seotud laiendid, mis sõltuvad otseselt verbi semantilisest tähendusest ning mille grammatilised ja

[2] Vt nt www.keeleveeb.ee (10.02.2010) ja www.cl.ut.ee (10.02.2010).

semantilised omadused tulenevad verbi argumentstruktuurist. Võtsime arvesse nii obligatoorsed kui fakultatiivsed seotud laiendid, verbi vabad laiendid püüdsime vaatluse alt välja jätta, välja arvatud juhtudel, kui nad olid lause algul ja mõjutasid otseselt lause kohustuslike elementide sõnajärge, nt ***Kord hommikupoolikul*** *helistas mulle sõber Joonas*. Nõnda jäid küll kõrvale mitmed õppijaile raskusi valmistavad lausetasandi sõnajärje probleemid (sh määruste omavaheline paiknemine), kuid see valik kitsendas pisut meie ees seisvat üsna ulatuslikku ülesannet. Kuna tekstides tuli sageli ette komplementlauseid, milleta pealause oleks poolik, siis võtsime arvesse ka mõned osalausekujulised verbilaiendid.

Käesolevas artiklis nimetame oma uuritavat struktuuri tuumlauseks. Lähenemine eesti keele sõnajärje korrektsuse hindamisele põhines kirjakeele korpusel kui emakeelekõnelejate poolt kirjapandud ja toimetatud tekstil, sõnajärje uurimustes esiletoodud seaduspärasustel ning autorite introspektsioonil. Meie koostatud õigete (paremate) sõnajärjemallide kogu esindab pigem tendentse kui reegleid, neist kõrvalekaldumist ei saa sageli veaks lugeda, sest neid võib ette tulla ka emakeelsete kõnelejate tekstides. Hinnates õppijakeele lauseid, otsustasime kahtluse korral lausejärje siiski tinglikult küsitavaks (ebasoovitavaks) hinnata, sest leidsime, et on parem, kui jääb alles võimalus nendele kohtadele tähelepanu juhtida.

Meie tööl oli kolm **eesmärki**:

1) püüda automaatselt tuvastada õppija tuumlausete struktuuri;
2) püüda automaatselt hinnata tuumlause argumentide jm elementide järje vastavust mallile (õigsust);
3) saada ülevaade teguritest, mis õppijakeele sõnajärje arvutianalüüsi positiivselt ja negatiivselt mõjutasid.

Võtsime korrektse eesti keele sõnajärje leidmisel näitematerjaliks Tartu Ülikooli kirjakeelekorpuse (TÜKK)[3] ilukirjanduse osa. Tegime õppijakeele sõnajärje vigade analüüsimisel selle valiku, sest ilukirjanduse sõnajärge on eesti keeles põhjalikumalt uuritud (nt Remmel 1963, Huumo 1994, Lindström 2005). Lisaks lausete tekstisidususele sobisid ilukirjandustekstid tänu narratiivsele ja arutlevale tekstiliigile (kasutasime enam jutustavaid ja võimalikult vähe dialoogi sisaldavaid tekste), inimese elu ja vaba aega puudutavale temaatikale ja hinnangute rohkusele. Ka vabakirjutuslikes õppijatekstides jutustatakse sageli autori elust ja kogemustest ning antakse mitmesugustele nähtustele hinnanguid. Ilukirjandustekstid erinevad õppijate tekstidest lausete pikkuse ja kohatise keerukusega: õppijakeele lausete hindamise juures ei ole oluliselt abi rohkete osalausete, täiendite ja verbilaiendite komplekssest analüüsist.

4. Õppija sõnajärje- ja tuumargumendivead

Kõigi vajalike ja õiges vormis tuumargumentide valimine lausesse ning nende õige järjestamine on edasijõudnud keeleõppija jaoks sageli selleks "koera sabaks", millest ei jõuagi üle astuda ning mis võibki jääda eristama õppija keelt emakeelekõneleja omast. Samas leidub EVKK-s loomulikult ka rohkesti näiteid kõigi soovituslike sõnajärjemallide kasutamisest sarnaselt emakeelekõnelejatele.

[3] http://www.cl.ut.ee/korpused/baaskorpus/ (10.02.2010).

Toome siinkohal valiku probleemidest, mis õppijail tekkisid, ning seletame neid Remmeli (1963) esitatud eesti keele sõnajärje seaduspärasuste alusel (vt ka Helle Metslang jt 2003: 118–121, 139–141, 149–151). Vajadusel lisame selgituseks praegu käibivad terminid "Eesti keele grammatika" (EKG II) mõistestikust. Siinse jaotise lausenäited pärinevad EVKK-st. Vealeidja edukust vaadeldud eksimustega toimetulekul vaatleme jaotises 6.

Õppijakeele lausetasandi sõnajärje kõrvalekalletest moodustas V2-malli mittejärgimine üle 2/3, näiteks: *Ta alati kuhugi* ***hilineb*** (parem: *Ta* ***hilineb*** *alati kuhugi*). Inversiooniga lauseid leidus õppijatekstides üldse alla 1% (mis on tunduvalt vähem kui emakeelsete kirjutajate tekstides, vrd Taela andmed jaotises 2). Õppijalausetes esines lisaks tavalisele V2-reegli vastu eksimisele järgmisi probleeme.

Üldjuhul põhjustab kõrvalekaldeid V2-st liigse(te) moodustaja(te) toomine predikaadi ette, kuid rohkelt leidub ka näiteid, kus üheks predikaadi ees paiknevaks lauseelemendiks on üldlaiend: *Vabal ajal Narvas ma käin discol* .. (parem: *Narvas käin ma vabal ajal diskol*); *Minu arvates see* ***on*** *inemeste tavad ja traditsioonid* (parem: *Minu arvates* ***on*** *need inimeste tavad ja traditsioonid*); *Aga samuti ma* ***olin*** *loomaaias* (parem: (*Aga*) *samuti* ***olin*** *ma loomaaias*).

Öeldistäite neutraalseim asukoht on lause lõpus (Remmel 1963: 261, 323). *See punkt on* ***väga tähtis*** *minu jaoks* (parem: *See punkt on minu jaoks* ***väga tähtis***). *Otsekohe tundus* ***teissugune*** *õhkkond* (parem: *Otsekohe tundus õhkkond* ***teistsugune***). Viimase näite ajamääruse õiget asukohavalikut kinnitavad järgmised seaduspärad. Pealause öeldise laiend võib Remmeli (1963: 258) järgi paikneda nii lause alguses, öeldise järel kui aluse ja öeldise vahel. Kui lauses on kaks öeldise laiendit, üks neist küsimusele *millal?* vastav ajamäärus, siis nende lahutamise korral saame kõige neutraalsema lause, kui algusse paigutame just ajamääruse (Remmel 1963: 227).

On verbe, mis väljendavad terviklikku mõtet ainult koos teise, ***ma-*** **või** ***da*****-infinitiivi vormis oleva verbiga. Seetõttu on neid verbe laiendaval infinitiivil tendents lause lõppu asetuda.** (Remmel 1963: 264) *Ning ei ole vaja* ***oodata*** *neid koolist* (parem: (*Ning*) *ei ole vaja neid koolist* ***oodata***).

Remmel on täheldanud, et **kui lauses on öeldistäitemäärus** (EKG II: latiivne seisundimäärus)**, siis asetub ta lõppu, sest ta annab lausele terviklikkuse.** Lisaks osutab ta, et **väliskohakäändes olev sihitismäärus** (EKG II: valdajamäärus või sõltuvusmäärus) **ei anna lausele terviklikkuse iseloomu ja asetub seepärast teistest öeldise laienditest ettepoole.** (Remmel 1963: 260) *Kõik muutused said* ***tõeliseks sokiks*** *inimestele* .. (parem: *Kõik muutused said inimestele* ***tõeliseks šokiks***).

Järgmine lause esindab kõrvalekallet eesti keeles väga sagedase *vaja olema*-lause tavajärjest: *Energia on vaja* ***inimestele*** (parem: *Energiat on* ***inimestele*** *vaja*). Remmel (1963: 238–239) **loeb kogejalause semantilise subjekti sihitismääruseks** (EKG II: valdajamäärus) **ning on leidnud, et see paikneb öeldise suhtes üldiselt samas asendis nagu tavalausete alus, s.t pealauses öeldisega kõrvuasendis.**

Kui personaalses konstruktsioonis on öeldiseks sihiline verb, mis selleks, et lause terviklik oleks, ei vaja muud laiendit peale sihitise, asub sihitis lause lõppu ja teised öeldise laiendid sellest ettepoole (Remmel 1963: 259). Järgmises õppijalauses, kus predikaadiks vaid üht seotud laiendit

nõudev verb, on sihitise asukoht lauses küsitav: *Ta pakkub **palju võimalusi** edaspidi* (parem: *Ta pakub edaspidi **palju võimalusi***)[4].

Keerulisem on põhjendada, miks mõjub järgmise lause järg küsitavalt: *Aga samuti reisimine annab **võimalusi** inimestele puhkama ja nautima loodusega* (parem: (*Aga*) *samuti annab reisimine inimestele **võimalusi** puhata ja loodust nautida*). Siin on probleemiks nimisõnalise peasõna ning järeltäiendi lahutamine. Kuid see selgitus ei anna vastust küsimusele, miks on parem tuua määrus sihitisest ettepoole – siin on vaja vaadata mitut Remmeli sõnastatud seadust koos. Remmel (1963: 259) toob välja tendentsi, et **lauses, kus peale sihitise on** (seotud laiendina – autorite kommentaar) **ka liikumise sihtkohta märkiv kohamäärus, asub viimane lõpus ja sihitis vahetult selle ees.** Samas toob ta näiteid, kuidas **sihitis liigub lõpu poole keerukama konstruktsiooniga lausetes, näiteks kui nimisõna laiend seisab oma põhisõna järel** (1963: 266, 327). Need reeglid näivad laienevat ka muudele sihti märkivatele määrustele, nt valdajamäärustele, ning võiksid siinset juhtumit selgitada[5].

Vaegisikulise konstruktsiooni kasutamisega kaasnev kõrvalekalle on: *Võib võrrelda **nende Big Ben'i** meie Oleviste kirikuga* (parem: ***Nende Big Beni** võib võrrelda meie Oleviste kirikuga*).

Tuumlause sõnajärje probleemist on lahutamatu tekstilause **kõigi vajalike tuumelementide olemasolu** küsimus lauses. Õppijatekstides tuleb ette nende ekslikku väljajätmist, mittevajalike elementide lisamist ning vale süntaktilise rolli kasutust (näiteks objekti asemel adverbiaal). Õppijate lausemoodustusraskusteks olid meie materjalis predikaadi puudumine (*Kõrval seina juures diivan*, parem: *Kõrval seina juures **on** diivan*); koopula puudumine (*See on ei ole suur tuba, aga päikesepaisteline*, parem: *See ei ole suur tuba, kuid **on** päikesepaisteline*); subjekti ja verbi väljajätt predikatiivilauses (*Väga raske küsimus*, antud kontekstis olnuks parem: ***See on** väga raske küsimus*); kohustusliku verbilaiendi puudumine (*Ma tahan alustada, et ..*, parem: *Ma tahan alustada **sellega**, et ..*); objekti kasutamine rektsiooniadverbiaali asemel (*Puhkuse lõpuks mina saanud aru **üks asi** ..*, parem: *Puhkuse lõpuks sain ma **ühest asjast** aru ..*); lisaks ka puuduv sidend komplementlause algul (*Aga ma arvan küll, kõige parem ja armas koht on ..*, parem: (*Aga*) *ma arvan küll, **et** kõige parem ja armsam koht on ..*).

5. Sõnajärje mallikogu ja vealeidja prototüübi loomine

Meetodeid, mida sõnajärje vealeidja programmi loomise aluseks valida, leidub palju, näiteks lingvistilistel reeglitel (vt nt Arppe 2000) ning statistikal põhinevad (Athanaselis jt 2006). Meie valikuks oli reeglipõhine lähenemine, et kasutada juba olemasolevaid eesti keele ressursse ja tarkvara. Võtsime vealeidja loomisel aluseks Kaili Müürisepa loodud kitsenduste grammatika süntaksianalüsaatori (edaspidi: parser)[6] ja morfosüntaktiliselt märgendatud ilukirjanduskorpuse[7]. Töö koosnes järgmistest etappidest: 1) esialgse mallide kogumi loomine ilukirjanduskorpuse ja

[4] Selgitus kehtib, kui lugeda võimalikuks, et *pakkuma* võib eesti keeles esineda nii mono- kui ditransitiivse verbina (vrd: *Ta pakub edaspidi palju võimalusi. Ta pakub **mulle** edaspidi palju võimalusi*).

[5] Pooli (1999) sõnastikus on ligi 106 sellist verbi, millest esialgsel vaatlusel 79 puhul mõjub sihti märkiva määrusega lõppev lause tavaliselt tõepoolest neutraalsemalt (*Kohandasin artikli **nõuetele***) ja vaid 27 juures on neutraalsem sihitisega lõppev lause (*Selgitasin sõbrale **oma ideed***). Vaatluse alt jäid välja kahe laiendiga verbid, millest üks oli sulgude abil vabaks laiendiks märgitud.

[6] EstCG Parser 1.0a, http://www.cs.ut.ee/~kaili/parser/, http://lepo.it.da.ut.ee/~kaili/grammatika/ (19.02.2010).

[7] http://www.cl.ut.ee/korpused/syntaksikorpus/ (11.02.2010).

EVKK väljavõtete põhjal, 2) õigete ja valede mallide automaatne lisamine kogusse ning käsitsi kontrollimine, 3) kogus olevate õigete mallide kasutatavuse hindamine kirjakeele ja õppijakeele korpuses, 4) vealeidja prototüübi loomine, mis annab mallikoguga võrdluse põhjal õppijalausete sõnajärjele hinnangu ja 5) vealeidja otsuste pisteline käsitsi kontroll ning tulemuste hindamine.

5.1. Sõnajärje mallid ja nende tuvastamise võimalused

Alustasime tööd emakeelekõnelejate lausetest pärinevate mallide kogumisega. Mallikogu koostamiseks oli vaja uurida lauseid nende formaalse esituse kujul, milleks sobis lausete esitamine morfosüntaktiliste märgendite abil. Parseri väljund kujutabki endast lause põhjalikku tõlgendust, kus on esitatud kõigi sõnade sõnaliik, vorm ning süntaktiline roll (parseri analüüs on pindsüntaktiline ning ei kajasta lause hierarhilist struktuuri). Formaalseks ning otsitavaks malliks peame siinkohal järjestatud komplekssete morfoloogiliste ja süntaktiliste tähiste (märgendite) jada, mis esitab lause struktuuri ning milles võivad esineda ainult vajalikud ehk malli eristamise seisukohalt olulised osad. Mallid koosnevad parseri väljundist valitud grammatilistest tähistest, nende selgitused on lisatud artikli lõppu.

Mallide otsimise etapil olid probleemiks traditsiooniliste arvutuslike meetodite kasutamise raskused. Olemasolevad tuntud keeleanalüüsi programmid on mõeldud korpusuuringuteks ja on kohandatud töötlema tekstimaterjali, mis on esitatud tähtede ja kirjavahemärkidega ega sisalda muid sümboleid. Kui sõnadest koosnevas tekstis saab niisugune tarkvara (näiteks WordSmith Tools 5.0) eraldada näiteks sõnadest koosnevaid kollokatsioone, siis formaalsete tähistega see kahjuks ei tööta. Loomuliku keele sõnalõpu tunnuseks on tühik või kirjavahemärk. Morfosüntaktilise märgendi lõpu tunnuseks ei pruugi tühik sobida, sest tühikud paiknevad ka märgendi sees. Näiteks on lauses *Mul on mälestused koolipõlvest* vormi *koolipõlvest* morfosüntaktiline esitus järgmine:

kooli_põli+st //_S_ com sg el // @<NN.

Kaalusime võimalust kasutada tuumargumentide järjestuse hindamisel *n*-grammmeetodit (vt nt Alam jt 2006) ning vastavat tarkvara.[8] Nägime siin raskuskohtadena süsteemi madalat töökiirust ning suurt varieeruvust õppijakeele ortograafias ja vormimoodustuses. Lisaks esitab vaid osa sõnale vastava morfosüntaktiliste tähiste kompleksi märgenditest tuumlause sõnajärje jaoks olulist informatsiooni. Need esinevad teiste, ebaoluliste märgendite vahel ning see struktuur ei sobi *n*-grammi moodustamiseks. On vaja algoritmi, mis lubaks selekteerida välja olulised märgendid ning uurida nende järjestust ning korduvust.

Eesmärgiks on leida suurim võimalik kogus nii õigeid kui ka küsitavaid (edaspidi tinglikult: valesid või ebakorrektseid) malle. Selline lähenemine annab võimaluse programmi sisendi täpsemaks liigitamiseks sõnajärje korrektsuse osas. Formaalsete mallide otsimiseks vajasime abivahendeid, mis tooksid esile korduvad grammatiliste märgendite kooslused. Neid kirjeldab jaotis 5.2.

[8] *n*-gramm on *n* järjest paikneva tähe või vajadusel sõna jada. Näiteks lausest *Poiss leidis võtmed üles* võib eraldada järgmised bigrammid (kahest osast koosnev jada): *poiss leidis*, *leidis võtmed*, *võtmed üles*. Korpusest leitud *n*-grammide sageduse põhjal on võimalik prognoosida, millise tõenäosusega ilmub vaadeldav osa, samuti millise tõenäosusega ilmub järgnev sõna, kui esimene sõna on fikseeritud. Meetodi rakendamiseks on võimalik kasutada spetsiaalset tarkvara, näiteks kfNgram: http://kwicfinder.com/kfNgram/kfNgramHelp.html (11.02.2010).

5.2. Mallileidja töö ettevalmistus

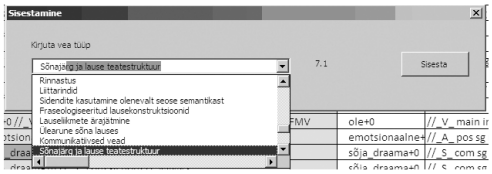
Kasutasime malliotsija prototüübi sisendina parseri väljundit. Õigete mallide otsimise juures kasutasime abivahendina tabelit, kus kirjakeelsete lausete järele olid ritta paigutatud kõigi lause sõnade morfosüntaktilised märgendid (ebakorrektsete märgendijadade otsimiseks kasutasime sarnast tabelit normist kõrvalekalduvaks märgitud õppijalausetest ning nende märgenditest). Kuid tabelite loomine eeldas parseri väljundi eelnevat kohandamist sobivale kujule.

Märgendijadade kogu loomiseks kopeeritakse eelnevalt parseri väljund etteantud tabelisse, mis on loodud Microsoft Exceli keskkonnas Visual Basic for Applications programmeerimisvahendite abil. Teksti sisestamiseks õigesse tabeli osasse on loodud spetsiaalne vorm (nupp "Sisesta tekst"). Seejärel on võimalik programmeeritud makrode abil teha parseri väljund korda nii, et sõnad, sõnade vormid ning morfosüntaktilised tähised oleksid eraldi tulpades (joonis 1).

	A	B	C	D	E	F	G	H	I	J	K	L
1	Sisesta tekst											
2												
3	Vea tüüp		Süntaksianalüsaatori väljund	Vea tüüp	Sõna vorm	Morfosüntaktiline analüüs	1	1.1		1.2	1.2.	1.2.
4	Tekst korda	Paiguta	LA				0	0	0	0	0	0
5	1	Koolist	kool+st //_S_ com sg el #cap // **CLB @ADVL		kool+st	//_S_ com sg el #cap // **CLB @ADVL	0	0	0	0	0	0
6	2	ma	mina+0 //_P_ pers ps1 sg nom // @SUBJ	7.1	mina+0	//_P_ pers ps1 sg nom // @SUBJ	0	0	0	0	0	0
7	3	ei	ei+0 //_V_ aux neg // @NEG		ei+0	//_V_ aux neg // @NEG	0	0	0	0	0	0
8	4	või	või+0 //_V_ mod indic pres ps neg #FinV #inf // @+FCV		või+0	//_V_ mod indic pres ps neg #FinV #inf // @+FCV	0	0	0	0	0	0
9	5	midagi	miski+dagi //_P_ indef sg part // @ADVL		miski+dagi	//_P_ indef sg part // @ADVL	0	0	0	0	0	0
10	6	eriti	eriti+0 //_D_ // @ADVL		eriti+0	//_D_ // @ADVL	0	0	0	0	0	0
11	7	halba	halb+0 //_A_ pos sg part // @<AN		halb+0	//_A_ pos sg part // @<AN	0	0	0	0	0	0
12	8	üttelda	üttelda+0 //_S_ com sg nom #? // @PRD		üttelda+0	//_S_ com sg nom #? // @PRD	0	0	0	0	0	0
13	9	$.	. //_Z_ Fst //		.	//_Z_ Fst //	0	0	0	0	0	0
14	10	LL					0	0	0	0	0	0
15	sisesta						0	0	0	0	0	0

Joonis 1. Mallide otsimise eeltöö. Parseri väljundi kohandamine Microsoft Exceli keskkonda

Kui tegu on õppijakeele lausetega, siis järgmise etapina on vaja tekst märgendada, et tuua esile ühe või teise veaga seotud kohti. Töö lihtsustamiseks on lisatud vorm, mis esimeste sisestatud tähtede järgi pakub veaklassifikaatori numbri (joonis 2; EVKK loomisel on vorm kasutusel ka teiste vealiikide märgendamisel). Õppijalausete korrektsust hindas keelespetsialist samas tabelis käsitsi. Valede mallide otsingul olid kasutusel vaid sõnajärje vea märgendi saanud laused.



Joonis 2. Vorm veaklassifikaatorite lisamiseks

Lause ning sellele vastavad töödeldud morfosüntaktilised tähised paigutatakse automaatselt eraldi Exceli töölehele (joonis 3).

	A	B	C	D
1	Ometi ma ei saanud oma vana s	//_D_ #cap // **CLB @ADVL	//_P_ pers ps1 sg nom // @SUBJ	//_V_ aux neg // @NEG
2	Smerdjakov oli Kesk-Venemaalt	//_S_ prop sg nom #cap #? // **CLB @SUBJ	//_V_ main indic impf ps3 sg ps af #FinV #Intr // @+FMV	//_S_ prop sg abl #cap // @ADVL
3	See on professionaalne vilumus	//_P_ dem sg nom #cap // **CLB @SUBJ	//_V_ main indic pres ps3 sg ps af #FinV #Intr // @+FMV	//_A_ pos sg nom // @AN>
4	Smerdjakov oli mulle mõjunud Š	//_S_ prop sg nom #cap #? // **CLB @SUBJ	//_V_ aux indic impf ps3 sg ps af #FinV #Intr // @+FCV	//_P_ pers ps1 sg all // @ADVL
5	Ta sarnanes saagiahnele õngem	//_P_ pers ps3 sg nom #cap // **CLB @SUBJ	//_V_ main indic impf ps3 sg ps af #FinV #Intr // @+FMV	//_A_ pos sg all // @AN>

Joonis 3. Mallide loomisele aluseks olevate morfosüntaktiliste märgendijadade tabel

Korrektsete lausete korral saab otsida, mis kombinatsioonid korduvad, ja selliste korduvate osade abil kirjutada malle õige lause tuvastamiseks. Sõnajärjevigadega lausetest leitud mallid lisasime valede mallide kogusse.

Mallide otsimiseks on vaja käsitsi moodustada a) tekstisõnade ning b) morfosüntaktiliste märgendite hulgad (kasutame terminit *hulk* siin matemaatilise hulga mõistes), millega sõnajärje uurimisel ning mallide tuvastamisel ei pea arvestama.

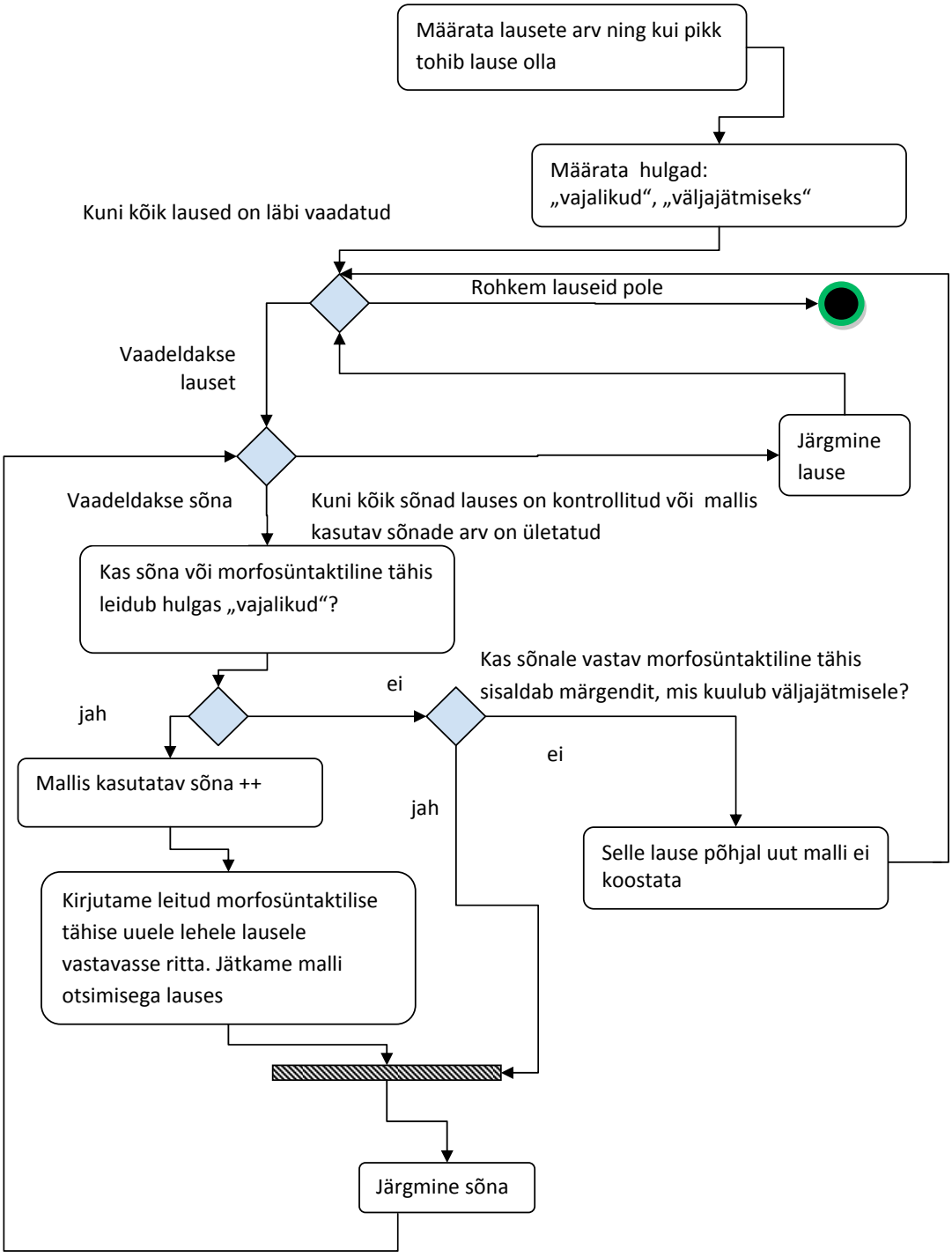
Fraasisisesed laiendid (moodustajatüüpidena nt omadussõnalised eestäiendid märgendiga @AN>, kvantori järellaiendid @<Q; loomuliku keele sõnaloenditena nt adjektiivi ja adverbi laiendid nagu *päris*, *väga*) jäävad vaatluse alt välja. Teises väljajäetavate lauseelementide hulgas olid üldlaiendina esinevad ühendid ja rõhumäärsõnad (nt *minu meelest*, *ju*, *siis*, *ka*, *hoopis*) ning sidesõnad, mis ei pruugi põhjustada subjekti ja predikaadi inversiooni ning millega on vaja seetõttu tavapärase V2-reegli puhul arvestada. Selle ning ka adjektiivi- ja adverbiaalifraasi laiendliikmete hulga reeglistik vajab edaspidi veel täpsustamist, sest nende sõnade käitumine on sõnajärje suhtes varieeruv ning nad võivad esineda ka erinevates funktsioonides (nt *siis* võib olla ka ajamääruse rollis). Lisaks on vaja moodustada hulk nendest märgenditest, mis on vajalikud ning mille abil on võimalik hinnata, kas lause sõnajärg on korrektne või mitte. Tuumlause sõnajärje analüüsi jäetakse sisse predikaat, ta argumendid (nt @SUBJ) ja vabad määrused.

5.3. Mallide otsimise algoritm

Mallide otsimiseks kontrollitakse märgendijadade tabelis kõigepealt lausete arvu ning arvutatakse maksimaalne lause pikkus vaadeldavas lausetekogumis. See informatsioon on vajalik, et tagada, et iga lause ning iga sõna saaks töödeldud. Järgnevalt loetakse sisse nii vajalike kui väljajäetavate sõnade hulgad. Eri tüüpi hulkade olemasolust tulenevalt peab lauset olema samaaegselt võimalik analüüsida nii loomuliku keele sõnajärjendi kui ka märgenditejada kujul (vt jaotis 5.2). Programmil tuleb analüüsida nii terviklikke lihtlauseid kui liitlause osalauseid (vt jaotis 3) – seetõttu lisasime algoritmile ka lihtsamad osalause piiri reeglid.

Mallide tuvastamise protsessi käigus, kuni kõiki lauseid ei ole veel läbi vaadatud, võtab programm uurimiseks järjekorras järgmise lause (vt joonis 4). Lausesiselt vaadeldakse sõnu üksteise järel. Kõigepealt otsitakse sõna väljajätmiseks lubatud loomuliku keele sõnade hulgast. Kui sõna sealt ei leita, siis kontrollitakse vajalike märgendite hulka. Kui sellesse kuuluv sobiv märgend leitakse, siis kirjutatakse märgend välja vastavasse lahtrisse ning minnakse järgmise sõna analüüsi juurde. Kui sõna või märgend leidub aga väljajäetavates hulkades, siis ei kirjutata seda vastavasse lahtrisse välja, vaid lihtsalt jätkatakse analüüsi, minnes järgmise sõna juurde. Juhul kui nimetatud hulkadest vastet ei leita, lause analüüs katkestatakse. Näiteks osalause piiri märgend CLB ei kuulu ülaltoodud hulkadesse ning kui tegu ei ole just lause vasakpoolseima osalause alguse märgendiga, siis selle lause uurimine lõpetatakse ning minnakse järgmise lause analüüsi juurde.

Tabel 1 esitab näitena lause *Juba paar nädalat valitses põud* (TÜKK) analüüsi.



Joonis 4. Mallide otsimise algoritm. Joonis näitab malliotsija otsuseid ühe lause ning ka kogu vaadeldava teksti töötlemisel. Plokkskeemil kujutatud protsessis kirjutatakse sobivate lausete põhjal mallikogusse uued mallid, mittesobivad laused jäetakse vahele. Rombidega esitatakse hargnemisi ja tsükleid, tehtega ++ näidatakse sammu (muutuja) suurenemist, programmi lõppu tähistatakse ringiga

Tabel 1. Mallileidja sisend

Juba	paar	nädalat	valitses	põud	.
//_D_ #cap // **CLB @ADVL	//_N_ card sg nom l // @ADVL	//_S_ com sg part // @<Q	//_V_ main indic impf ps3 sg ps af #FinV #Part-P // @FMV	//_S_ com sg nom // @SUBJ	//_Z_ Fst //

Sõna *juba* leitakse loomuliku keele sõnade hulgast, mis on lubatud välja jätta, seda välja ei kirjutata ning minnakse järgmise sõna juurde. Järgmine sõna *paar* ei asu loomuliku keele sõnade hulgas ning tema morfosüntaktiline info ei sisalda ka märgendit, mis oleks lubatud väljajätmiseks. Seejuures aga leitakse märgend @ADVL, mis on sõnajärje seisukohalt oluline ning asub hulgas "vajalikud". See märgend kirjutatakse vastavasse lahtrisse välja, seejärel minnakse järgmise sõna juurde. Sõna *nädalat* on lubatud väljajätmiseks, sest selle märgend @<Q kuulub vastavasse hulka. Märgendit ei kirjutata välja ning vaatluse alla võetakse järgmine sõna. Selleks on *valitses*, mille morfosüntaktilises analüüsis leidub märgend @FMV, mis on pärit hulgast "vajalikud". Märgend kirjutatakse välja ning asutakse uurima sõna *põud*. Märgend @SUBJ on samuti hulgas "vajalikud" ning kirjutatakse välja. Järgmisel kohal asub punkt ning sellele vastav märgend, millega analüüs lõpetatakse, kuna see ei kuulu ühtegi hulka. Tulemuseks saame malli: ['@ADVL' '@FMV' '@SUBJ'].

Eelkirjeldatud algoritmi abil saadud mallid läbisid keelespetsialisti kontrolli, mille käigus täiendati hulkasid ja täpsustati märgendeid mallides.

5.4. Õigete mallide kogu

Ilukirjanduskorpuse tekstide põhjal eraldati 148 malli, neile lisati lingvisti poolt käsitsi moodustatud 94 malli. Viimaste eesmärk on parandada süsteemi suutlikkust tulla toime eri tüüpi ühend- ja väljendverbe sisaldavate lausetega ning subjektiellipsiga lausetega.

Mallikogus domineerivad lihtsad süntaktilised rollid (nt @SUBJ), kuid sõnu on ka piiritletud näiteks ainult sõnaliigi (ühendverbi koosseisu kuuluv adverb) või käände abil (kogejalause esimese sõna vormiks on adessiiv või allatiiv). Tabel 2 iseloomustab meie mallikogu lähemalt.

Tabel 2. Lauseliikmete positsioonieelistused mallikogus (242 malli)[9]

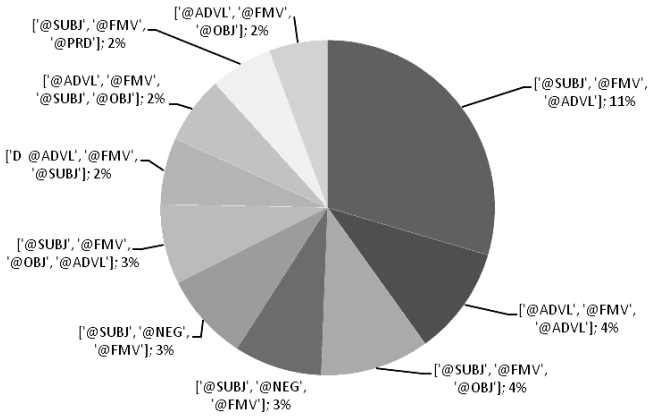
Lauseliige	Esinemus kokku	Positsioon lauses					
		1.	2.	3.	4.	5.	6.
@SUBJ	130	88	4	19	15	4	–
@OBJ	133	1	18	37	53	20	4
@PRD	10	1	2	4	1	2	–
@ADVL	147	29	6	57	35	17	3
@FCV	63	12	48	3	–	–	–
@FMV	95	16	59	15	5	–	–
@IMV	82	–	12	21	29	20	–

Mallide sees esinevad esimesel positsioonil kõige sagedamini subjekt (88 korda) ning adverbiaal (29). Malli teisel positsioonil on ootuspäraselt populaarseimad lauseliikmed finiitsed predikaadid ning mitmeosaliste predikaatide finiitsed osad (vastavalt @FMV 59 korda ning @FCV 48 korda). Malli kolmandal positsioonil domineerivad adverbiaal (57, põhiliselt abimäärsõnad) ja objekt (37). Predikatiiv esineb kõige rohkem kolmandal kohal (4 korda). Tabeli alusel on edaspidi võimalik

[9] Mallikogus jagunevad adverbiaalid kaheks: adverbist peasõnaga määrused (eelkõige ühendverbide komponendina) ning määrused, mille fraasi peasõna sõnaliiki pole täpsustatud. Adverbist peasõnaga määrused eelistavad 3. ja 4. positsiooni, täpsustamata peasõnaga adverbiaalid esinevad põhiliselt 1. positsioonil.

moodustada ja süsteemiga liita reegliteblokk lauseliikmete positsioonidest, mis on suure tõenäosusega ebasoovitavad, nt subjekt teisel kohal ning mitmeosalise predikaadi infiniitvorm lause esimesel positsioonil.

242 malli leidmise järel kontrolliti suurema ilukirjanduse valimiga (681 lauset) nende esinemissagedust. Joonis 5 näitab protsentuaalset jaotust sagedasemate mallide vahel (mida esines 681 lause seas üle 1%). Uuritud valimis osutusid kõige populaarsemateks järjestusteks kolmikud ['@SUBJ', '@FMV', '@ADVL'] (73 lauset ehk 681-st mallist 11%, nt *Smerdjakov tuli ujumast*) ning ['@ADVL', '@FMV', '@ADVL'] (26 lauset ehk 4%, nt *Pärast istusime sauna ees terrassil* – mallide esinemissageduse hindamisel ja vealeidja töös lubasime malle sobitada lausega ka siis, kui adverbiaalide arv täpselt ei klappinud – vt lähemalt jaotis 7). Enamik sageli kasutatavatest mallidest ongi kolmekohalised. Seejuures on aga populaarsed ka mõned neljakohalised mallid: ['@SUBJ', '@FMV', '@OBJ', '@ADVL'], mille esinemissagedus on 3% (19 lauset, nt *Ta juhtis meie tähelepanu pääsukestele*) ning ['@ADVL', '@FMV', '@SUBJ', '@OBJ'] (*Seepeale võttis ohvitser relva*). 2%-line esinemissagedus on ka määrsõnalise adverbiaaliga (sageli üldlaiendiga) algaval mallil ['_D_ .. @ADVL', '@FMV', '@SUBJ'] (16 lauset, nt *Samas aga kõlasid lasud*). Suurem osa malle olid aga haruldased, esinedes vaid 1–2 korda. Seetõttu ei anna vealeidja arendamisel mallidehulga kasvatamine enam olulist efekti. Tulemuste edasiseks parandamiseks tasub täiustada mallide lausetega ühendamise viise.

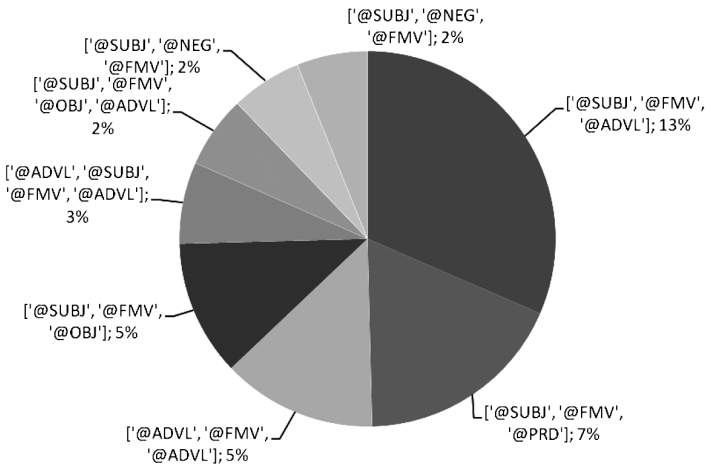


Joonis 5. Sõnajärjemallide esinemus kirjakeeles (681 TÜ ilukirjanduskorpuse lause seas)

5.5. Kirjakeele mallide esinemus õppijakeelekorpuses

Võrdlesime kirjakeele korpusest leitud õigete mallide kogu õppijakeelega. Selleks võtsime õppijakeele korpusest juhumeetodil välja 4743 lausest koosneva valimi. Üldjoontes on tulemused võrreldavad ilukirjanduse omadega, kuid nad on ebatäpsemad lausetes esinevate õppijavigade ja mõningate keerukamate süntaksiküsimuste tõttu, mille lahendamine malle otsivale programmile hetkel üle jõu käib (vt jaotis 7). Toome siinkohal kokkuvõtte õppijate eelistustest tuumlause sõnajärjele.

Materjalis leiduvad õppijakeele lausestruktuurid vastasid vaid osale 242-st ilukirjanduse põhjal leitud mallist: õppija kasutas neist vaid 83, mis katsid ligi 3/4 õppijakeele lausetest.



Joonis 6. Õppijate eelistused sõnajärjemallide suhtes

Populaarsemate mallide osas oli nii kattuvusi kui erinevusi. Sarnaselt ilukirjanduskeelele oli õppijakeele valimis kõige sagedamini kasutavaks malliks ['@SUBJ', '@FMV', '@ADVL'], mille osakaal on 13% (564 lauset, nt *Zaura õpib kohalikus koolis*). Teisel kohal on aga mall ['@SUBJ', '@FMV', '@PRD'] osakaaluga 7% (324 lauset, näiteks *Aga arvutite kasutamine on ka ohtlik*), mis ilukirjanduses oli ainult 2% osakaaluga. Kolmandal kohal on mall ['@ADVL', '@FMV', '@ADVL'] (5%, 238 lauset, nt *Esialgu olime paanikas*).

5.6. Õppijakeele sõnajärje vealeidja prototüüp

Jaotistes 5.4 ja 5.5 iseloomustasime sõnajärje mallide kogu, mis on sõnajärje vealeidjale sisendmooduliks. Järgnevalt anname ülevaate praeguseks väljatöötatud esimesest vealeidja prototüübist, mis on mõeldud rakendamiseks EVKK õppijakeelekorpuses.

Vealeidja eesmärk on anda automaatselt hinnang õppijateksti lausetele (vaid antud projektis vaadeldavatele, mitte kõigile – vt jaotis 3), mis on programmi sisestatud: kas nende tuumelementide sõnajärg on korrektne või suure tõenäosusega ebasoovitav.

Prototüüp kasutab Unixi-põhist kitsenduste grammatika süntaksianalüsaatorit (parserit), mis on installeeritud korpusega seotud serverisse. Vealeidja hinnang põhineb õppijalausete võrdlusel õigete sõnajärjemallide koguga (tulevikus ka valede sõnajärjemallide koguga ning lauseliikmete tüüpiliste positsioonieelistuste infoga). Programmi töö koosneb järgmistest sammudest:

1) morfosüntaktiliste tähiste jadade lisamine õppijalausetele parseri abil;
2) õppijalause sõnavormide ning morfosüntaktiliste tähiste võrdlemine mittevajalike sõnavormide ja märgendite hulkadega; neisse kuuluvate sõnade kustutamine vaatlusalustest järjenditest;
3) allesjäänud morfosüntaktiliste tähiste seast hulka “vajalikud” kuuluvate märgendite esiletõstmine (üle jäävad (osa)lausepiiri ja kirjavahemärkide märgendid);
4) esiletõstetud märgendite jadade võrdlemine õigete mallide koguga, võttes arvesse mallide rakendamise prioriteete (näiteks kui sama lause jaoks on

tuvastatud kaks malli ning üks neist koosneb kolmest ning teine neljast osast, siis kasutatakse pikemat malli). Valede mallide kogu käesoleva katse käigus ei rakendatud;

5) malliga ühendatud lause puhul malli katvuse hindamine lause piires (nt 100% või 75%); malliga vaid vähesel määral kaetud lausete küsitavaks märkimine;
6) malliga mitteseostatud lausete küsitavaks märkimine.

6. Vealeidja tulemused

Vealeidja töö hindamiseks kontrollisime 100 õppijalauset, mida programm oli analüüsinud süntaksianalüsaatoriga ja ühendanud olemasoleva malliga või tuvastanud vastava malli puudumise mallikogust. 100 lause seas oli 26 tuumlause sõnajärje veaga lauset. Neist 16 puhul tuvastas programm, et lause ei vasta mitte ühelegi sõnajärje mallile. 8 puhul ühendas programm lause olemasoleva malliga ekslikult. 100 lausest 26 puhul viis programm korrektse lause kokku õige sõnajärje malliga ja 36 puhul tuvastas õigesti vastava malli puudumise mallikogus. Lausete võrdlemisel mallikoguga tegi programm 21 juhul vea õppijavea tõttu, 4 korral eelkõige süntaksianalüsaatori eksliku analüüsi tõttu ning 10 korral keerukamate grammatikanähtuste esinemisel, millega ta toime ei tulnud. 3 lauset jäid analüüsist välja, kuna need ei kuulunud vealeidja vaatlusaluste lausetüüpide alla, kuigi vealeidjale on ette antud ka reeglid eemaldamaks analüüsist mitte-vaatlusalused laused: näiteks need, mis sisaldavad impersonaali ja algavad võrdlusega.

Parseri ekslikke analüüse ja seeläbi ka vealeidja valetõlgendusi põhjustasid järgmised õppija vead:

- ortograafiaviga (***Koguaeg*** *ma olin maganud ainult kolm või neli korda olin ärkanud*; *Muidugi ka Eesti rahvas mängis isesesvumisprotsessis* ***surt*** *osa*);
- ortograafiaviga koos objektikäände veaga (***Stokholmis*** *olen näinud väga huvilise* ***asi***);
- ajavormi viga (*Mul* ***olnud*** *vähe aega, et hästi uurida seda kohta*; *Puhkuse lõpuks mina* ***saanud*** *aru üks asi, et ma olen linna elanik ja eluga külas kohanematu*);
- ekslik infinitiivivormi valik (*Meie ajal tehnikaareng annab palju võimalusi* ***reisima***);
- objektikäände viga (*Edasi ma olen näinud surepärane* ***linn***; *Oma* ***matk*** *ma alustasin veekogust*);
- ühildumisviga (*Seal mulle* ***meeldis lõvid****, sest ma olen näinud nende esimene kord elus*);
- rektsiooniviga (ekslik adverbiaali kasutamine objekti asemel: *Veel ma meenutan kultuursetest* ***inimestest***);
- osalausepiiri märkimise viga (*Õues jalutasid koduloomad mina esimene kord silmasin lehm ei ole televisorist*);
- osalausepiiri märkimise viga ja predikaadi ekslik puudumine (*Estlased väga toredad inimesed mina tutvustan ja suhtlen nendega hea meelega*).

Parser suutis hoolimata vahelejäetud komast osalause piiri mitmel korral tuvastada, kui õppija oli seal kasutanud sidendit (*Sellepärast ma ei saa midagi öelda mida ma olen näinud esmaspäeval*).

Lisaks põhjustasid süntaksianalüsaatori ekslikke analüüse subjektita kogejalause (*Kõigepealt mulle meeldis selles linnas*) ja tundmatu lühend.

7. Vabad ja seotud adverbiaalid

Vealeidja saab parseri abil õppijakeelega sageli hästi hakkama, leides lausest viiteid ka tundmatute või ebakorrektsete kohtade analüüsiks. Keerukamad probleemid õigete tuumargumentide olemasolu ja nende järjestuse tuvastamisel tekkisid nii õppijakeele kui kirjakeele tekste analüüsides ennekõike kahel juhul: komplekssemate rinnastusjuhtudega (nt *varased rongile tõttajad ja hilja peale jäänud pidulised*) ning arvukate vabade adverbiaalidega lausete korral. Nimetatud adverbiaale on raske automaatselt ära tunda ning nad segavad lausete kokkuviimist õigete mallidega.

Kõige teravamaks probleemiks oligi seotud adverbiaalide eristamine vabadest. Kui seotud laiend on nii sisult kui ka vormilt põhjast tingitud, siis vaba laiendi tingitus põhjast on minimaalne: laiend võib esineda täistähenduslikult väga erinevate põhjadega. Püüdsime mallide koostamisel vabu adverbiaale vältida, sest nende arv on potentsiaalselt lõpmatu. Verbilaiendite seotuse üle otsustamine võib lisaks sõltuda ka subjektiivsest keeletajust. Seotud laiendite äratundmiseks on vaja teada eri verbide rektsioone, mis võiks lisaks aidata hinnata ka sõnajärje õigsust: Huno Rätsepa (1978: 218–221) järgi on vähemalt paljude kontekstiväliste lausete puhul kehtiv, et sõnajärje tüübi valiku määrab verb (vrd *lendama* ja *leiduma*: *Lind lendab*; *Leidus vabatahtlikke*).

Kuna oluline osa kohustuslikest verbilaienditest kajastub Raili Pooli sõnastikus "Eesti keele verbirektsioone" (objektid ja seotud adverbiaalid) ning eesti keele verbikesksete püsiühendite andmebaasis (ühend- ja väljendverbide osad)[10], võiks adverbiaalide seotuse üle otsustamisel olla nende allikate kasutamisest abi. Katse käigus läbivaadatud 50 õppijalauses esines 108 adverbiaali, millest 70 olid vabad adverbiaalid või üldlaiendid (neist 27 lausealgulised ning seetõttu üldjuhul mallidesse kuuluvad)[11], 8 abimäärsõnad (7 neist on olemas ka eesti keele verbikesksete püsiühendite andmebaasis), 25 seotud adverbiaalid (17 neist on olemas Pooli sõnastikus), 2 adverbiaalatribuudid ning 3 relatiivlause alguse sidendid.

Kirjeldame lähemalt Pooli "Eesti keele verbirektsioone" (1999) kasutamise võimalusi adverbiaalide automaatanalüüsil. Sõnastik esitab rektsioonid, sh objektikäänded 563 verbile ning võimaldab lihtsat tekstilausete kohustuslike elementide tuvastamist. Verbirektsioonide sõnastiku kasutamise väärtuseks on suure hulga vabade adverbiaalide analüüsist kõrvaldamine, mida muude formaalsete tunnuste abil ei saa teha. Kuigi vabade laiendite eraldamine seotutest polnud ilmselt raamatu autori eesmärk, toetab see spetsiaalsete vahendite puudumisel robustset tuvastust. Järgnevalt kirjeldame näidet, kuidas sõnastikku "Eesti keele verbirektsioone" on võimalik selleks kasutada.

Korrektselt moodustatud lause sisaldab malli järgi lause alguse adverbiaali või adverbiaale (kui on), üldjuhul subjekti, verbi, sõnastiku verbikirjes toodud verbilaiendit ning malli kõiki muid vajalikke elemente, mis on leitud viisil, nagu käesolevas artik-

[10] http://www.cl.ut.ee/ressursid/pysiyhendid/index.php?lang=et (2.01.2010).

[11] Parseri märgenduspõhimõtetest lähtuvalt on loetud adverbiaalide hulka ka üldlaiendid.

lis kirjeldatud. Muude elementide hulka ei tohiks kuuluda adverbiaalid, sest liigsete vabade adverbiaalide eemaldamine on üks sõnastiku kasutamise eesmärkidest.

Näiteks on sõnastikus verbi *ahvatlema* juures kirjed:

objekt + millele?	(*Poiss ahvatles sõbra vargusele*)
objekt + *ma*-infinitiiv	(*Reklaam ahvatleb inimesi ostma*)

Lähtudes sõnastiku kirjetest, peab vealeidja verbi kohustuslikeks laienditeks üksköik mis käändes objekti ning allatiivis määrust (nt *vargusele*) või *ma*-infinitiivi (nt *ostma*). Sõnastiku kasutamine võimaldab aga lausest *S ahvatles L-i parki koristama šampanjaga*[12] eemaldada adverbiaali *šampanjaga* kui vaba laiendi, sest komitatiivset verbilaiendit verbikirjes pole. Samuti on näiteks õppijakeele lause *Ma tahan alustada, et igal suvel ma armastan kuhugi reisida* (EVKK) puhul võimalik tänu sõnastikule tuvastada lause tuumelemendi (komitatiivne määrus verbi *alustama* juures) puudumise viga.

Sõnajärje vealeidja peab arvestama ka võimalusega, et lause alguses võib esineda vabu adverbiaale, sest ka need mõjutavad lause tuumelementide sõnajärge. Näiteks: ***Sinisinise veikleva veepiiri taga*** *ahvatles vanderselle valge liivaranna ja männimetsaga väike püsielaniketa Pedassaar*.

Kui ühendada programmiga Pooli sõnastik, valmistaksid endiselt raskusi sõnastikust puuduvad seotud laiendid. Muud probleemid, mida sõnastiku kasutamine lahendada ei pruugi, on järgmised.

- Lause algul on rohkem kui üks vaba adverbiaal. Vahel on selline sõnajärg õige (***Ühe meremiili kaugusel sinisinise veikleva veepiiri taga*** *ahvatles vanderselle valge liivaranna ja männimetsaga väike püsielaniketa Pedassaar*), vahel aga mõjub ebaloomulikuna (***Vabal ajal Narvas*** *ma käin discol* .. (EVKK)).
- Valik ühe verbi kirjete vahel nõuab aeg-ajalt semantilist infot, mis on potentsiaalsete vealeidja vigade allikas (*arvestama + kellega? mida?* – elusolendit tähistav laiend ei saa üldjuhul olla partitiivis).
- Mitmesugused spetsiifilisemad konstruktsioonid, näiteks verbiga *olema* (*Oli* ***kuldilus*** *suvise pööripäeva järgne päev* – atribuut on kohustuslik, kuid sõnastik selliseid konstruktsioone ei kirjelda).

8. Kokkuvõte

Keeleõppija suutlikkus moodustada õigesti sihtkeele tuumlauset on vajalik kogu lause terviktähenduse edukaks edasiandmiseks ja sellistegi vigade vältimiseks, mida teevad kõige kõrgemal tasemel teise keele kõnelejad. Artiklis on kirjeldatud eesti keele kui teise keele õppija raskusi lause tuumelementide valikul, vormistamisel ja järjestamisel, antud ülevaade esimestest sammudest õppijakeele sõnajärje arvutianalüüsil, tutvustatud esimest sõnajärje vealeidja prototüüpi ja analüüsitud tegureid, mis selle tööd mõjutavad.

Vealeidja prototüübi väljatöötamisel loodi Tartu Ülikooli kirjakeelekorpuse põhjal sõnajärje mallide kogu, kasutades selleks süntaksianalüsaatorit ja Exceli keskkonnas loodud vahendeid. Tutvustatud vealeidja eesmärk on tuvastada õppijalause tuumargumendid, nende süntaktiline roll ja järjestuse õigsus. Vealeidja

[12] 7. jaotise näited pärinevad internetist, kui ei ole märgitud teisiti.

annab õppijalausete sõnajärje korrektsusele hinnangu, lähtudes peamiselt nende morfosüntaktiliste analüüside vastavusest mallidele.

Kirjeldatud katses sisestati vealeidjasse 100 EVKK lauset ning kontrolliti programmi väljundi õigsust. 62 korral tegi programm õige otsuse: 26 korral viis vealeidja lause kokku õige malliga ja tuvastas lause korrektsuse, 36 korral tuvastas vealeidja, et lause ei vasta mallile. Neist 36 lausest sisaldasid 16 õppijapoolset kõrvalekallet loomulikust sõnajärjest.

Lahendamist vajavad märkimata osalause piiride, keerukamate rinnastusjuhtumite ning seotud adverbiaalide tuvastamise küsimused. Seotud adverbiaalide eristamiseks vabadest adverbiaalidest pakkusime võimaliku lahendusena Pooli sõnastiku "Eesti keele verbirektsioone" ja eesti verbikesksete püsiühendite andmebaasi kasutamist. Edaspidi oleks õppija sõnajärje arvutianalüüsil kasu näiteks lähtumisest kõrvalekaldumistest Remmeli sõnajärjereeglitest ning kirjakeele laseliikmete positsioonieelistustest. Et lause sõnajärge efektiivsemalt kontrollida, saab programmi lisada tüüpiliste vigade mooduli. Sagedasemate vigade tuvastamine enne lausete kontrollimist mallikogust tõstab programmi töö täpsust ja kiirust. See on tõenäoliselt vähem töömahukas lahendus, kui oleks võtta kasutusele pragmaatiline info, mis on tegelikult eesti keele sõnajärje peamiseks mõjutajaks.

Lühendid ja sümbolid[13]

@ADVL	adverbiaal	_A_	adjektiiv
@AN>	adjektiiv eestäiendina	_D_	adverb
@FCV	mitmeosalise predikaadi finiitne osa (*olema* liitaegades jms)	_P_	pronoomen
		_S_	substantiiv
		_V_	verb
@FMV	finiitne predikaat	CLB	osalause piir
@IMV	infiniitne predikaat	V	finiitverb
@OBJ	objekt	V2,	
@PRD	predikatiiv	V2-reegel	*verb teisel kohal* -reegel
@SUBJ, S	subjekt	X	verbi laiend, lause kõrvalliige, mida võib olla üks või enam
@<NN	nimisõna järeltäiendina		
@<Q	kvantori järellaiend		

Viidatud kirjandus

Alam, Md. Jahangir; UzZaman, Naushad; Khan, Mumit 2006. N-gram based statistical grammar checker for Bangla and English. – Proceedings of 9th International Conference on Computer and Information Technology (ICCIT 2006), Dhaka, Bangladesh. http://www.panl10n.net/english/outputs/Working%20Papers/Bangladesh/Microsoft%20Word%20-%2019_N_112.pdf (28.02.2010).

Arppe, Antti 2000. Developing a grammar checker for Swedish. – Torbjorn Nordgard (Ed.). Proceedings from the 12th Nordiske datalingvistikkdager, Trondheim, December 9-10, 1999. Department of Linguistics, Norwegian University of Science and Technology (NTNU). Trondheim: University of Trondheim, 13–27. http://www.ling.helsinki.fi/~aarppe/Publications/Nodalida-99.pdf (02.01.2010).

Athanaselis, Theologos; Bakamidis, Stelios; Dologlou, Ioannis 2006. A fast algorithm for words reordering based on language mode. – 16th International Conference, Athens,

[13] Loendis on toodud vaid artiklis sagedamini kasutatud märgendid. Täielik kitsenduste grammatika märgendite ülevaade asub veebiaadressidel http://math.ut.ee/~kaili/parser/demo/morftags.html ja http://math.ut.ee/~kaili/papers/syntax.html (12.02.2010).

Greece, September 10-14, 2006. Proceedings, Part II, 943–951. http://www.springerlink.com/content/q646768285871122/fulltext.pdf (02.01.2010).
EKG II = Erelt, Mati; Kasik, Reet; Metslang, Helle; Rajandi, Henno; Ross, Kristiina; Saari, Henn; Tael, Kaja; Vare, Silvi 1993. Eesti keele grammatika II. Süntaks. Lisa: Kiri. Eesti Teaduste Akadeemia Keele ja Kirjanduse Instituut. Tallinn.
Huumo, Tuomas 1994. Kontrastiivinen tutkimus suomen ja viron sanajärjestyksestä. Litsensiaaditöö. Käsikiri Turu Ülikooli soome ja üldkeeleteaduse õppetoolis.
Höglund, Kirsi 2006. Estnisk satsföljd och meningsstruktur. Doktoriväitekiri. Uppsala.
Lindström, Liina 2005. Finiitverbi asend lauses. Dissertationes philologiae estonicae Universitatis Tartuensis, 16. Tartu: Tartu Ülikooli kirjastus.
Metslang, Helle; Krall, Ingrid; Pajusalu, Renate; Saarso, Kristi; Sõrmus, Elle; Vare, Silvi 2003. Keelehärm: eesti keele probleemseid piirkondi. Tallinn: Tallinna Pedagoogikaülikool.
Pool, Raili 1999. Eesti keele verbirektsioone. Tartu: Tartu Ülikooli kirjastus.
Remmel, Nikolai 1963. Sõnajärjestus eesti lauses. – Eesti keele süntaksi küsimusi. KKI uurimused, VIII. Tallinn: Eesti Riiklik Kirjastus, 216–381.
Rätsep, Huno 1978. Eesti keele lihtlausete tüübid. Emakeele Seltsi toimetised, 12. Tallinn: Valgus.
Tael, Kaja 1988. Sõnajärjemallid eesti keeles (võrrelduna soome keelega). Preprint KKI-56. Tallinn.

Võrgumaterjalid

Eesti keele verbikesksete püsiühendite andmebaas. http://www.cl.ut.ee/ressursid/pysiyhendid/index.php?lang=et (2.01.2010).
EstCG Parser 1.0a. http://www.cs.ut.ee/~kaili/parser/, http://lepo.it.da.ut.ee/~kaili/grammatika/ (19.02.2010).
EstCG Parseri morfoloogilised märgendid. http://math.ut.ee/~kaili/parser/demo/morftags.html (12.02.2010).
EstCG Parseri süntaktilised märgendid. http://math.ut.ee/~kaili/papers/syntax.html (12.02.2010).
EVKK = Eesti vahekeele korpus. evkk.tlu.ee (20.10.2009).
Keeleveeb. www.keeleveeb.ee (10.02.2010).
kfNgram. http://kwicfinder.com/kfNgram/kfNgramHelp.html (11.02.2010).
Süntaktiliselt analüüsitud ja ühestatud tekstikorpus. http://www.cl.ut.ee/korpused/syntaksikorpus/ (10.02.2010).
Tartu Ülikooli Arvutilingvistika Uurimisrühma koduleht. www.cl.ut.ee (10.02.2010).
TÜKK = Tartu Ülikooli kirjakeele korpus. http://www.cl.ut.ee/korpused/baaskorpus/ www.cl.ut.ee (20.10.2009).

Helena Metslangi (Tallinna Ülikool, Tartu Ülikool) teaduslikeks huvialadeks on süntaks, eriti lauseliikmete vahelised üleminekualad, ja õppijakeele korpused.
helena.metslang@gmail.com

Erika Matsaki (Tallinna Ülikool) uurimisvaldkonnad on loogilised konstruktsioonid eestikeelsetes tekstides, keeletehnoloogia.
erika.matsak@tlu.ee

AUTOMATIC WORD ORDER ANALYSIS OF ESTONIAN AS A SECOND LANGUAGE: THE NUCLEAR SENTENCE

Helena Metslang, Erika Matsak
Tallinn University, University of Tartu

This article gives an overview of our work on the automatic analysis of second language word order. For this purpose, an error analyzer and a set of correct word order patterns found from the fiction sub-corpus of Tartu University's Corpus of Written Estonian were created. It is important to be able to form the nuclear sentence of the target language well (incl. subject, finite verb, obligatory modifiers of the verb and other elements influencing the sentence word order) because a well-formed core clause conveys the integral meaning of the whole sentence and helps to avoid the errors that even the very high level learners make.

The article describes the learner's difficulties in choosing, inflecting and ordering the core elements of the sentence (in the data of EVKK – Estonian Interlanguage Corpus). It gives an overview of the first steps of the automatic analysis of learner language word order, introduces the set of correct word order patterns and the prototype of the word order error analyzer and analyzes the factors influencing the success of its performance.

The task of the error analyzer is to detect the core arguments, estimate their syntactic role and assess the correctness of their order. In the test described in the article 100 sentences were analyzed with the error analyzer and the output was assessed. The program made the correct choice 62 times of which 26 times the error analyzer connected the correct sentence with the right rule and assessed that the word order in the clause was correct. In the other 36 times the analyzer found that the input clause didn't correspond to any rule in the pre-defined set. 16 clauses from these 36 contained a word order error.

There are a number of problems that still need to be solved including the erringly unmarked clausal border, more complex cases of coordination and imprecise adverbial analysis. The article suggests the use of the valency dictionary of Estonian verbs (Pool 1999) and the Database of Estonian verbal multi-word expressions as a possible solution of how to improve the program's ability to distinguish free and bound adverbials. In the future it would be useful to integrate Remmel's (1963) word order laws and the typical position preferences of parts of sentences into the system. To check the learners' word order more effectively it is possible to add the module of typical errors to the programme. Identifying the more frequent errors of the clause, before searching within the set of correct patterns, would raise the precision and speed of the programme. This would likely be less time consuming than supplementing the error analyzer with pragmatic information (which in fact is one of the most important factors influencing Estonian word order).

Keywords: word order, corpus linguistics, second language acquisition, Estonian